REPORT

Alignment Analysis of Secondary Mathematics Standards and the SAT Reasoning Test

Maine

Norman L. Webb

April 10, 2006

REPORT

Alignment Analysis of Secondary Mathematics Standards and the SAT Reasoning Test

Maine

Norman L. Webb

April 10, 2006

REPORT

Alignment Analysis of Secondary Mathematics Standards and the SAT Reasoning Test Maine

Norman L. Webb

April 10, 2006

This study is one of two alignment studies conducted for the State of Maine. An Alignment Analysis Institute was held December 8, 2005, in Madison, Wisconsin, to analyze the secondary mathematics standards and the mathematics SAT Reasoning Test. This study was done to provide an external analysis of a previous study conducted by the College Board and reported in September, 2005. This report consists of a description of the four criteria used to judge the alignment between Maine Content Standards for mathematics for high school and one form of the SAT Reasoning Test. This report includes tables listing the results of two reviewers' coding of the assessments and standards.

REPORT

Alignment Analysis of Secondary Mathematics Standards and the SAT Reasoning Test

Maine

Norman L. Webb

December 31, 2005

REPORT

Alignment Analysis of Secondary Mathematics Standards and the SAT Reasoning Test Maine

Norman L. Webb

December 31, 2005

This study is one of two alignment studies conducted for the State of Maine. An Alignment Analysis Institute was held December 8, 2005, in Madison, Wisconsin, to analyze the secondary mathematics standards and the mathematics SAT Reasoning Test. This study was done to provide an external analysis of a previous study conducted by the College Board and reported in September, 2005. This report consists of a description of the four criteria used to judge the alignment between Maine Content Standards for mathematics for high school and one form of the SAT Reasoning Test. This report includes tables listing the results of two reviewers' coding of the assessments and standards.

Acknowledgements

Reviewers:

Norman Webb (Group Leader)	WI
Rob Ely	WI

The State of Maine funded this analysis. Horace (Brud) Maxcy and Valerie Seaberg, Coordinators of the Maine Educational Assessment, from the Maine Department of Education were the main contacts for the Department and oversaw the coordination of the study.

Table of Contents

Executive Summary

An alignment analysis was conducted on December 8, 2005, in Madison, Wisconsin. The agreement of the Maine Learning Results in Mathematics for high school and the SAT Reasoning Test were analyzed. Two reviewers with extensive content expertise and knowledge of the process conducted the analysis. Results from this study are compared to a study conducted by the College Board of the same set of standards and test.

The analysis indicated that major improvements would be required for the 54-item SAT Reasoning Test and the mathematics Maine Learning Results to be fully aligned. The assessment was judged to have only one or two items corresponding to six of the 11 standards, too few to make a judgment on students' performances related to these standards. For the five standards that had an adequate number of items, three of these did not fully meet the Depth-of-Knowledge Correspondence criterion. That is, too high of a proportion of the DOK levels of the assessment items were lower than the DOK level of the corresponding performance indicators. The three and four standards that did not fully meet the range and balance criterion were those with a low number of corresponding items so these alignment criteria were not a major issue.

The two analyses, one by two Wisconsin reviewers and the other by the College Board, produced similar results in the proportion of items assigned to specific standards. However, the College Board reviewers assigned a DOK level 2 (skills and concepts) to a higher proportion of the performance indicators than did the Wisconsin reviewers. As a consequence, the Wisconsin analysis indicated more of an issue with the DOK levels of the items. The College Board analysis found no alignment issues with regards to DOK levels of the items. From the results of the Wisconsin analysis, nearly 40 items would need to be replaced or added to achieve full alignment between the SAT and the Maine Learning Results. This number could be reduced if the items were robust enough to measure content related to more than one performance indicator and standard. Algebra and geometry are over emphasized on the SAT, beyond the requirements used in this analysis to determine an acceptable alignment level. Items related to these standards would be likely candidates for replacement by items measuring content related to standards like discrete mathematics and probability that only have one or two items.

Alignment Analysis of Secondary Mathematics Standards and the SAT Reasoning Test
Maine

Norman L. Webb

Introduction

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute of any one of these two system components. Alignment describes the match between expectations and assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Language Arts and Science Education* (Webb, 1997).

A one-day Alignment Analysis Institute was conducted December 8, 2005, in Madison, Wisconsin. Two mathematics education content experts who are experienced alignment reviewers and who have taught high school or college mathematics served as the reviewers. The Maine mathematics standards for secondary education were compared to a released form of the SAT Reasoning Test. The same form of the assessment was used in this analysis as was used in a study conducted by the College Board, reported in September, 2005 (College Board, 2005).

For the purposes of this analysis, we have employed the convention of standards and objectives to describe two levels of expectations for what students are to know and do. Standard as used here refers to the Maine Learning Results secondary (grades 9–12) content standards. Each of the 11 mathematics standards (A through K) is comprised of one to five performance indicators, or objectives. It is assumed that the performance indicators or objectives are intended to span the content of the standards under which they fall. The standards and objectives are reproduced in Appendix A.

The two reviewers were well familiar with the alignment coding process. They reviewed the procedures at the beginning of the analysis, but did not receive any formal training. The two reviewers did go over the definitions of the four depth-of-knowledge (DOK) levels. Then the reviewers participated in 1) a consensus process to determine the depth-of-knowledge levels of the Maine content objectives and 2) individual analyses of the assessment items of each of the assessments.

To derive the results on the degree of agreement between the Maine mathematics standards and the SAT, the reviewers' responses were averaged. Any variance among reviewers is considered legitimate, with the true depth-of-knowledge level for the item falling somewhere between two or more assigned values. Such variation could signify a lack of clarity in how the objectives were written, the robustness of an item that can legitimately correspond to more than one objective, and/or a depth of knowledge that falls in between two of the four defined levels. Reviewers were allowed to identify one assessment item as corresponding to up to three objectives—one primary hit (objective) and up to two secondary hits. Reviewers were instructed to use multiple hits for one item if appropriate. Reviewers could only code one depth-of-knowledge level to each assessment item, even if the item corresponded to more than one objective.

Reviewers were instructed to focus primarily on the alignment between the state standards and the SAT. However, they were encouraged to offer their opinions on the quality of the standards, or of the assessment activities/items, by writing a note about the item. Reviewers could also indicate whether there was a source-of-challenge issue with the item—i.e., a problem with the item that might cause the student who knows the material to give a wrong answer, or enable someone who does not have the knowledge being tested to answer the item correctly. For example, a mathematics item that involves an excessive amount of reading may represent a source-of-challenge issue because the skill required to answer is more a reading skill than a mathematics skill. Reviewers did not identify any item with a source-of-challenge issue. However, reviewers did write several notes clarifying their rationale for their coding. In many cases, reviewers' notes referenced a difficulty in finding a precise match between an assessment item and a performance indicator.

The results produced from the institute pertain only to the issue of agreement between the Maine state standards and the SAT Reasoning Test. Note that this alignment analysis does not serve as external verification of the general quality of the state's standards or the SAT. Rather, only the degree of alignment is discussed in these results. The averages of the reviewers' coding were used to determine whether the alignment criteria were met. When reviewers did vary in their judgments, the averages lessened the error that might result from any one reviewer's finding. Standard deviations are reported, which give one indication of the variance among reviewers.

To report on the results of an alignment study of Maine's Learning Results and the SAT, the study addressed specific criteria related to the content agreement between the state standards and grade-level assessments. Four alignment criteria received major attention: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation.

Alignment Criteria Used for This Analysis

This analysis, which judged the alignment between standards and assessments on the basis of four criteria, also reported on the quality of items by identifying items with sources of challenge and other issues. For each alignment criterion, an acceptable level was defined by what would be required to assure that a student had met the standards.

Categorical Concurrence

An important aspect of alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents. This criterion was judged by determining whether the assessment included items measuring content from each standard. The analysis assumed that the assessment had to have at least six items measuring content from a standard in order for an acceptable level of categorical concurrence to exist between the standard and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors have to be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63. This indicates that about 63% of the group would be consistently classified as masters or nonmasters if two equivalent test administrations were employed. The agreement coefficient would increase if the cutoff score were increased to one standard deviation from the mean to .77 and, with a cutoff score of 1.5 standard deviations from the mean, to .88. Usually, states do not report student results by standards, or require students to achieve a specified cutoff score on subscales related to a standard. If a state did do this, then the state would seek a higher agreement coefficient than .63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a standard and as a basis for making some decisions about students' knowledge of that standard. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

Depth-of-Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the standard, as judged in this analysis, at least 50% of the items corresponding to an objective had to be at or above the level of knowledge of the objective: 50%, a conservative cutoff point, is based on the assumption that a minimal passing score for any one standard of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the corresponding objectives. For example, assume an assessment included six items related to one standard and students were required to answer correctly four of those items to be judged proficient—i.e., 67% of the items. If three, 50%, of the six items were at or above the depth-of-knowledge level of the corresponding objectives, then to achieve a proficient score would require the student to answer correctly at least one item at or above the depth-of-knowledge level of one objective. Some leeway was used in the analysis on this criterion. If a standard had between 40% and 50% of items at or above the depth-of-knowledge levels of the objectives, then it was reported that the criterion was “weakly” met.

Interpreting and assigning depth-of-knowledge levels to both objectives within standards and assessment items is an essential requirement of alignment analysis. These descriptions help to clarify what the different levels represent in mathematics:

Level 1 (Recall) includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify a Level 1 include “identify,” “recall,” “recognize,” “use,” and “measure.” Verbs such as “describe” and “explain” could be classified at different levels, depending on what is to be described and explained.

Level 2 (Skill/Concept) includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include “classify,” “organize,” ”estimate,” “make observations,” “collect and display data,” and “compare data.” These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as “explain,” “describe,” or “interpret,” could be classified at different levels depending on the object of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, also is a Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Level 2 activities are not limited to just number skills, but can involve visualization skills and probability skills. Other Level 2 activities include noticing and describing non-trivial patterns, explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3.
Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

Level 4 (Extended Thinking) requires complex reasoning, planning, developing, and thinking, most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as at Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be at Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include developing and proving conjectures; designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Range-of-Knowledge Correspondence

For standards and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of objectives within the standard with one related assessment item/activity. Fifty percent of the objectives for a standard had to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a standard. This assumes that each objective for a standard should be given equal weight. Depending on the balance in the distribution of items and the necessity for having a low number of items related to any one objective, the requirement that assessment items need to be related to more than 50% of the objectives for a standard increases the likelihood that students will have to demonstrate knowledge on more than one objective per standard to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the objectives. However, any restriction on the number of items included on the test will place an upper limit on the number of objectives that can be assessed. Range-of-knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of standards and a large number of objectives. If 50% or more of the objectives for a standard had a corresponding assessment item, then the range-of-knowledge criterion was met. If between 40% and 50% of the objectives for a standard had a corresponding assessment item, the criterion was "weakly" met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range-of-knowledge criterion only considers the number of objectives within a standard hit (a standard with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among these objectives. *The balance-of-representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another.* An index is used to judge the distribution of assessment items. This index only considers the objectives for a standard that have at least one hit—i.e., one related assessment item per objective. The index is computed by considering the difference in the proportion of objectives and the proportion of hits assigned to the objective. An index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a standard are equally distributed among the objectives for the given standard. Index values that approach 0 signify that a large proportion of the hits are on only one or two of all of the objectives hit. Depending on the number of objectives and the number of hits, a unimodal distribution (most items related to one objective and only one item related to each of the remaining objectives) has an index value of less than .5. A bimodal distribution has an index value of around .55 or .6. Index values of .7 or higher indicate that items/activities are distributed among all of the objectives at least to some degree (e.g., every objective has at least two items) and is used as the acceptable level on this criterion. Index values between .6 and .7 indicate the balance-of-representation criterion has only been "weakly" met.

Source-of-Challenge

The source-of-challenge criterion is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted mathematics skill, concept, or application. Cultural bias or specialized knowledge could be reasons for an item to have a source-of-challenge problem. Such item characteristics may result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed.

Findings

Standards

Two reviewers participated in the depth-of-knowledge (DOK) level consensus process for the standards and performance indicators for the Maine mathematics standards. A summary of their deliberations is presented in Table 1. The complete group consensus values for each standard and objective can be found in Appendix A. It should be noted that the two reviewers' assignment of the DOK levels differ some from the DOK levels assigned in the College Board analysis. Overall, the two Wisconsin reviewers assigned a higher percentage of performance indicators to a DOK level 3 than did the College Board reviewers. The Wisconsin reviewers judged that 35% of the mathematics performance indicators had a DOK level 3. The College Board reviewers indicated that 19% of the performance indicators had a DOK level 3 and one (3%) had a DOK level 4. The two groups of reviewers differed on a total of 11 of the 31 performance indicators (35%). The greatest difference between the two groups was in assigning DOK levels to performance indicators under Standard C (Data Analysis and Statistics) and under Standard G (Patterns, Relations, and Functions). These differences will not necessarily produce differences in the attainment of the alignment criteria as long as reviewers consistently applied the DOK levels in judging both the performance indicators and the assessment items.

In this analysis, the complexity of the performance indicators was reasonable for the high school level with nearly all of the indicators with a DOK level of 2 or higher and over one-third with a DOK level 3.

The reviewers were told that within each of the 11 standards, the performance indicators were intended to fully span the content of that standard and, in turn, each goal is spanned by the performance indicators that fall under it. For this reason, the reviewers only coded items to a standard if there were no performance indicator that the item appeared to target. As indicated in Table 2, both reviewers coded 10 of the 54 items (18%) to a standard rather than a performance indicator. This is a high percentage and indicates that nearly one-fourth of the items on the mathematics SAT measured mathematical content not included in the performance indicators. This could be because the content was expected to be learned in an earlier grade or because the content is a higher level than Maine expects all students to know. The reviewers' notes (Table 9.7 in Appendix B) sometimes indicate the reason why an item was coded to a generic objective. For example, both reviewers noted that the reason they assigned item 49 to Standard D rather than D.1 was because the item did not have students determine the probability for a compound event, only a simple event.

Table 1

Percent of Objectives by Depth-of-Knowledge (DOK) Levels for Secondary Standards, Maine Alignment Analysis for Mathematics

Grade	Total number of objectives	DOK Level	# of objs by Level	% within std by Level
A. NUMBERS AND SENSE	2	2	2	100
B. COMPUTATION	2	2 3	1 1	50 50
C. DATA ANALYSIS AND STATISTICS	5	2 3	2 3	40 60
D. PROBABILITY	2	2 3	1 1	50 50
E. GEOMETRY	3	1 3	1 2	33 66
F. MEASUREMENT	2	1 2	1 1	50 50
G. PATTERNS, RELATIONS, FUNCTIONS	4	1 2 3	1 1 2	25 25 50
H. ALGEBRA CONCEPTS	4	2 3	3 1	75 25
I. DISCRETE MATHEMATICS	4	2	4	100
J. MATHEMATICAL REASONING	1	3	1	100
K. MATHEMATICAL COMMUNICATION	2	1 2	1 1	50 50
Total	31	1 2 3	4 16 11	12 51 35

Table 2

Items Coded to Generic Objectives by More Than One Reviewer, Maine Alignment Analysis for Mathematics with the SAT Reasoning Test

Grade	Assessment Item	Generic Objective (Number of Reviewers)
9-12	4,19,22,35	A
	12,23,39	B
	49	D
	51	G
	21	J

Alignment of Curriculum Standards and Assessments

The results from the alignment analysis are presented in Tables 3. “Yes” indicates that an acceptable level on the criterion was fully met. “WEAK” indicates that the criterion was nearly met, within a margin that could simply be due to error in the system. “NO” indicates that the criterion was not met by a noticeable margin. More detailed data on each of the criteria are given in Appendix B in the first three tables for each of the grade levels. The first table in Appendix B, Table 9.1, lists the average number of items coded by the two reviewers for each standard.

Reviewers could code an item as measuring content related to more than one performance indicator. Reviewers were instructed to assign each item to a primary indicator. If the item produced information about a student’s knowledge of more than one indicator, then the reviewer could code an item to up to two secondary indicators. Reviewers used, on an average, 11 secondary hits in this analysis. The SAT assessment had 54 items. The two reviewers recorded, on the average, 65 hits. This is fewer than the 91 hits recorded in the analysis by the College Board reviewers. The Wisconsin reviewers coded about one-third of the items with secondary hits, indicating that the item measured content related to more than one performance indicator. The College Board reviewers coded nearly two-thirds of the items with a secondary hit.

This analysis indicates that the Maine Learning Results for mathematics and the SAT Reasoning Tests are not well aligned (Table 3). The assessment was found to be only fully aligned on all four alignment criteria used in this analysis with Standard H (Algebra Concepts). The assessment is partially or nearly aligned with four more of the 11 standards—Standard A (Numbers and Sense), Standard B (Computation), Standard E (Geometry), and Standard G (Patterns, Relations, and Functions). For these four standards only minor changes in the assessment or additions to the assessment would be needed to achieve full alignment. On the remaining six standards, the SAT Reasoning Test is not aligned mainly because the assessment includes too few items (only one or two items) to make a judgment about students’ performance related to these standards. This finding coincides with the analysis done by the College Board with one exception. The College Board analysis indicated that the assessment had eight items that corresponded to Standard K (Mathematical Communication). Only one Wisconsin reviewer coded one item as corresponding to this standard. The other reviewer found no items. The alignment is discussed in more detail by alignment categories.

Categorical Concurrence

The Wisconsin analysis and the College Board analysis had good agreement on the distribution of items among the 11 standards (Table 4). This was true even though the College Board reviewers used about twice the number of secondary hits. The Wisconsin analysis found that the assessment and standards had an acceptable level on the Categorical Concurrence criterion of six or more items for four standards (B, E, G, and A). For Standard A (Numbers and Sense) one Wisconsin reviewer assigned six items to this standard whereas the second reviewer only assigned five items. Therefore, the average hits for Standard A was 5.5. With only two reviewers, this is close enough to be considered as having met the acceptable level for Categorical Concurrence. The big difference between the two analyses was the assignment of items to Standard K (Mathematical Communication). The College Board reviewers assigned eight items as measuring content related to the first objective under Standard K (restate, create, and use definitions in mathematics to express understanding, classify figures, and determine the truth of a proposition or argument). One Wisconsin reviewer coded one item to this objective. The Wisconsin reviewers viewed that this objective requires students to verbally express or perform some action to be met. One reviewer judged that item 53

Table 3
Summary of Acceptable Levels on the Four Alignment Criteria for Maine Mathematics Standards and SAT Reasoning Test

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
Part 1: All Items Equal Weight				
A. NUMBERS AND SENSE	NO	YES	YES	YES
B. COMPUTATION	YES	NO	YES	YES
C. DATA ANALYSIS AND STATISTICS	NO	NO	NO	NO
D. PROBABILITY	NO	NO	NO	YES
E. GEOMETRY	YES	WEAK	YES	NO
F. MEASUREMENT	NO	YES	YES	YES
G. PATTERNS, RELATIONS, FUNCTIONS	YES	NO	YES	YES
H. ALGEBRA CONCEPTS	YES	YES	YES	YES
I. DISCRETE MATHEMATICS	NO	YES	NO	YES
J. MATHEMATICAL REASONING	NO	NO	YES	YES
K. MATHEMATICAL COMMUNICATION	NO	YES	NO	NO

(Section 8, item 15) did meet this requirement because the item asked about the truth of mathematical statements (inequalities). However, the mathematical statements were not presented as a proposition or argument. Other items did ask students to select what was true, but these items were coded by the Wisconsin reviewers to the mathematical topic under which students would study the definition (e.g. geometry) rather than under Standard K.

Depth-of-Knowledge Consistency

The biggest difference between the two analyses was in the proportion of the standards that had an acceptable level on the Depth-of-Knowledge Consistency criterion of 50% or more of the items with the same or higher DOK level when compared to the level of the corresponding objective. The College Board analysis indicated that the DOK criterion was met by all 11 standards. The Wisconsin analysis indicated that the criterion was only met by five of the 11 standards (A, F, H, I, and K) and weakly met by one other standard (E). Three of the six standards that did not fully meet the acceptable level for the DOK criterion by the Wisconsin analysis only had a very few number of items assigned to the standard (C, D, and J). For these standards, all of the items were judged to have a DOK level lower than the DOK level of the corresponding performance indicator.

For Standard B (Computations) the Wisconsin reviewers coded nearly all of the eight items, on the average, to the generic objective or standard. The standard was given a DOK level 3 because one of the two performance indicators was judged to have a DOK level 3. The performance indicator B.1 expects students to justify their results. However, all of the items mapped to Standard B were assigned by the reviewers a DOK level 2 mainly because they did not require students to justify their answers.

The Depth-of-Knowledge Consistency criterion was only weakly met for Standard E by the Wisconsin Analysis. Most of the items mapped to Standard E were judged to correspond to performance indicator E.2 (use inductive and deductive reasoning). This performance indicator was assigned a DOK level 3, but nearly all of the corresponding items had a DOK level 2. The two Wisconsin reviewers both agreed that item 33 was the only item that mapped to E.2 and had a DOK level 3.

On the average, the two Wisconsin reviewers found less than one item with an appropriate DOK level that was mapped to Standard G. The first two indicators—G.1 (create a graph to represent a real-life situation and draw inferences from it) and G.2 (translate and solve a real-life problem using symbolic language)—both were assigned a DOK level 3. However, all of the items on the assessment had a DOK level 2. Thus, the DOK criterion was not met for Standard G.

Most of the items on the mathematics SAT Reasoning Test were judged by the two Wisconsin Reviewers to have a DOK level 2. Only four of the items were rated by at least one of the two reviewers with a DOK level 3. However, over one-third of the performance indicators were judged to have a DOK level 3. The College Board reviewers rated a higher proportion of the objectives at a DOK level 2. For example, for Standard G

(Patterns, Relations, Functions), the College Board reviewers assigned all four of the performance indicators with a DOK level 2 whereas the Wisconsin reviewers assigned one with a DOK level 1, one with a DOK level 2, and two with a DOK level 3. The difference in the judgment of the complexity of the performance indicator is one explanation for the differences in the results on the Depth-of-Knowledge Consistency between the two analyses.

Table 4
Number and Percentage of Hits by Standard for Each Analysis for Mathematics

Standards	WI Analysis		College Board Analysis	
	Hits	Percent	Hits	Percent
A. NUMBERS AND SENSE	6	9	6	7
B. COMPUTATION	8	12	12	13
C. DATA ANALYSIS AND STATISTICS	2	3	3	3
D. PROBABILITY	1	1	1	1
E. GEOMETRY	12	18	8	9
F. MEASUREMENT	1	1	4	4
G. PATTERNS, RELATIONS, FUNCTIONS	8	12	13	14
H. ALGEBRA CONCEPTS	24	36	28	31
I. DISCRETE MATHEMATICS	2	3	5	5
J. MATHEMATICAL REASONING	2	3	3	3
K. MATHEMATICAL COMMUNICATION	1	1	8	9
Total	67*	100	91	100

* Decimal values were rounded to nearest whole number.

Range-of-Knowledge Correspondence

The low number of performance indictors makes it easier to satisfy an acceptable level on the Range-of-Knowledge Correspondence criterion. An acceptable level of at least 50% of the performance indicators with a corresponding item for the range criterion was met by seven of the standards. The range criterion was not met for the four standards that had only one or two corresponding items (Standards C, D, I, and K). This alignment issue is primarily the result of having too few items on the SAT that corresponded to these standards. The Range-of-Knowledge Correspondence criterion was weakly met by the same number of standards in the College Board analysis, but the standards differed some (Standards A, B, D, and K). In general, both analyses indicated that the range was acceptable when there were an adequate number of items.

Balance of Representation

In the Wisconsin analysis, the Balance of Representation criterion was not met by three of the 11 standards. For Standards C and K, this can be explained by a low number of assigned items with only one reviewer mapping items to these standards. For Standard G (Geometry), the imbalance is more due to the low number of reviewers rather than an over emphasis of one performance indicator compared to other performance indicators. Therefore, balance is not a major concern. In the College Board analysis, the balance criterion was met for all 11 standards.

Action Required to Achieve Full Alignment

In order for the SAT Reasoning Test in mathematics to be fully aligned with the Maine Learning Results for high school a number of items would need to be replaced or the test would need to be supplemented with additional items. The SAT assessment used in this analysis had nearly 24 algebra items that all mapped to Standard H and about 12 items that mapped to Standard E (Geometry). Although these are dominant high school mathematics topics, some of these items would be likely candidates for replacement. However, it would be impossible to meet the Categorical Concurrence requirement of six items for each standard with a 54-item assessment and 11 standards unless the items measured content related to more than one of the standards. If items that are replaced or added are carefully selected to have an appropriate DOK level and to target performance indicators not currently assessed, then full alignment could be achieved by these modifications:

Standard A	No action needed
Standard B	Replace four items with items having a DOK level 3
Standard C	Add four items
Standard D	Add five items
Standard E	Replace one or two items with those having a higher DOK level
Standard F	Add five items
Standard G	Replace four items with those having a higher DOK level
Standard H	No action needed
Standard I	Add five items
Standard J	Add four items
Standard K	Add six items

Source of Challenge

Reviewers were asked to indicate whether there was a source-of-challenge issue on any of the items. The concerns expressed by the reviewers are given in the fifth table

(Table 9.5) in Appendix B. Neither of the two reviewers identified any source-of-challenge issues.

Notes

The two reviewers made other comments about the items, which they recorded as notes. These notes are presented in the seventh table (Table 9.7) in Appendix B. Reviewers' notes sometimes clarify the match between the item and the objective as being weak. The notes also indicate issues that a reviewer might have found with an item and his/her suggestion regarding how the item could be improved.

General Comments made by Reviewers

After coding the assessment, the two reviewers together responded to four questions about their opinions of the general alignment between the standards and the assessments:

A. For each standard, did the items cover the most important topics you expected from the standard? If not, what topics were not assessed that should have been?
B. For each standard, did the items cover the most important performance (DOK levels) you expected of the standard? If not, what performance was not assessed?
C. Was there any content you expected to be assessed, but found no items assessing that content? What was that content?
D. What is your general opinion of the alignment between the standards and assessment:
   i. Perfect alignment
   ii. Acceptable alignment
   iii. Needs slight improvement
   iv. Needs major improvement
   v. Not aligned in any way.
E. Other Comments.

Their responses indicate the reflections of reviewers at the time of coding. They complement and inform the more rigorous analysis, but should not be interpreted as definitive, only impressionistic. The responses by the mathematics reviewers are presented below.

A. For each standard, did the items cover the most important topics you expected by the standard? If not, what topics were not assessed that should have been?

R1. Not really. The assessment items tended to correspond to one or two performance indicators under a standard rather than the full set of performance indicators. I expected to see more data analysis and more on function. There was nearly nothing on discrete mathematics.

Re standard B, the assessment included items requiring computation, but the items did not target the two objectives under computation—approximate the solutions and non-base 10 number systems.

There was very little under measurement. When there was an item related to area, I mainly coded this as E.2 (determine properties) because usually the item did not require a direct application of a formula.

R2. Do you mean "did the items cover the most important topics I expected," or "Maine expected" because Maine asks for things that aren't here. For instance, computation means approximation for Maine, and no items required this. Likewise, Standards G & H ask for real-life contexts, which the test doesn't approach. There were no items addressing compound events or distributions under probability.

B. For each standard, did the items cover the most important performance (DOK levels) you expected by the standard? If not, what performance was not assessed?

R1. The items required reasoning. Students who are able to answer the items correctly will have to know a range of mathematics at a conceptual and reasoning level. With respect to function, the test did not ask students to model phenomena with functions or identify a variety of situations. Instead, students were asked if a graph represented a description of a situation. The test did have some good items that required students to apply reasoning in determining the area of a figure (E.2).

Items did not require students to compute the probability of a compound event.

Under algebra, there were no items that required students to use tables to interpret expressions or equations and where students had to analyze a situation.

R2. Not entirely. For example, lots of geometry items, but there were no inductive reasoning as asked for in E2. No justification/reasonableness under B. No inferences or modeling under G.

C. Was there any content you expected to be assessed, but found no items assessing that content? What was that content?

R1. There were no items on statistical design as expected in C.3, C.4, and C.5.

R2. There were no data analysis items, or pattern recognition, matrices, networks, complex numbers, real number structure, trig, or geometric transformations.

D. What is your general opinion of the alignment between the standards and assessment?

R1. Needs major improvement to be fully alignment to the Maine Standards

R2. Needs major improvement

E. *Other comments:* (None)

Reliability Among Reviewers

The pairwise agreement among the two mathematics reviewers' assignment of DOK levels to items was .80. This is very high for two reviewers and shows strong agreement in assigning items a DOK level. One reason of this high level of agreement was there was very little variation in the DOK level of the items. Nearly all of the items had a DOK level of 2.

The pairwise agreement among the two reviewers in assigning items to standards was .62. In general, with eight or more reviewers, an agreement of .9 is desired. The pairwise agreement among the two reviewers in assigning items to performance indicators was .46, a little lower than desired. The lack of agreement among reviewers in assigning items to standards and performance indicators can be due in part to an overlap in the Maine Learning Results standards. For example, both Standard G and Standard H include performance indicators relating to graphs. Because the items with graphs did not precisely match performance indicators under either standard, the reviewers did vary in how the graphing items were mapped. Another reason for the lower agreement between the reviewers in assigning items to standards and performance indicators is because the items did not closely fit the wording in the performance indicators. As a consequence reviewers had to make more of a judgment about what was a fit leading to more variation between the reviewers Also, the reviewers only used the performance indicator statements and did not benefit from a deeper knowledge of the Maine Learning Results. The results for this analysis are computed by averaging results between the two reviewers. This helps to lessen the error or inconsistency among reviewers.

Summary

Major improvements are needed for the 54-item SAT Reasoning Test and the mathematics Maine Learning Results to be fully aligned. The assessment was judged to have only one or two items corresponding to six of the 11 standards, too few to make a judgment on students' performances related to these standards. For the five standards that had an adequate number of items, three of these did not fully meet the Depth-of-Knowledge Correspondence criterion indicating that too high of a proportion of the DOK levels for these items were lower than the DOK level of the corresponding performance indicators. The three and four standards that did not fully meet the range and balance criterion were those with a low number of corresponding items so these alignment criteria and are not considered a major issue.

The two analyses, one by two Wisconsin reviewers and the other by the College Board, produced similar results in the proportion of items assigned to specific standards. However, the College Board reviewers assigned a higher proportion of the performance indicators a DOK level 2 (skills and concepts) than did the Wisconsin reviewers. As a consequence, the Wisconsin analysis indicated more of an issue with the DOK levels of the items. The College Board analysis found no alignment issues with regards to DOK levels of the items.

From the results of the Wisconsin analysis, nearly 40 items would need to be replaced or added to achieve full alignment between the SAT and the Maine Learning Results. This number could be reduced if the new items are robust enough to measure content related to more than one performance indicator and standard. Algebra and geometry are over emphasized on the SAT, beyond what is needed for the assessment to be aligned with the standards. Items related to these standards would be likely candidates for replacement by items measuring content related to standards like discrete mathematics and probability that only have one or two items.

Two reviewers conducted this analysis. Normally, a full alignment analysis would require from six to eight reviewers. The two reviewers had reasonable agreement in assigning DOK levels to items. However, their agreement in assigning items to standards and performance was lower than desired. The final results were determined by averaging the results from each of the reviewers to lessen the error due to any one reviewer. The reviewers used statements of the standards and performance indicators, but did not have available to them other materials or access to people from the state with a greater understanding of what is included under a performance indicator. However, the reviewers, who are mathematics education content experts with extensive experience in interpreting performance indicators, assigned items to the performance indicators based on a common interpretation of their wording.

References

College Board. (2005). *Report for the State of Maine on the Alignment of the SAT and PSAT/NMSQT to the Maine Learning Results.* http://www.maine.gov/education/sat_initiative/docs/sat_alignment_study.pdf

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, *25*(1), 47-55.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in language arts and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Appendix A

Maine Mathematics Standards and Pair Consensus DOK Values

Table 13

Group Consensus

Maine Mathematics, Mathematics, High School

Level	Description	DOK
A.	NUMBERS AND SENSE: Students will understand and demonstrate a sense of what numbers mean and how they are used.	2
A.1.	Describe the structure of the real number system and identify its appropriate applications and limitations.	2
A.2.	Explain what complex numbers (real and imaginary) mean and describe some of their many uses.	2
B.	COMPUTATION: Students will understand and demonstrate computation skills.	3
B.1.	Use various techniques to approximate solutions, determine the reasonableness of answers, and justify the results.	3
B.2.	Explain operations with number systems other than base ten.	2
C.	DATA ANALYSIS AND STATISTICS: Students will understand and apply concepts of data analysis.	3
C.1.	Determine and evaluate the effect of variables on the results of data collection.	3
C.2.	Predict and draw conclusions from charts, tables, and graphs that summarize data from practical situations.	3
C.3.	Demonstrate an understanding of concepts of standard deviation and correlation and how they relate to data analysis.	2
C.4.	Demonstrate an understanding of the idea of random sampling and recognition of its role in statistical claims and designs for data collection.	2
C.5.	Revise studies to improve their validity (e.g., in terms of better sampling, better controls, or better data analysis techniques).	3
D.	PROBABILITY: Students will understand and apply concepts of probability.	3
D.1.	Find the probability of compound events and make predictions by applying probability theory.	2
D.2.	Create and interpret probability distributions.	3
E.	GEOMETRY: Students will understand and apply concepts from geometry.	3
E.1.	Draw coordinate representations of geometric figures and their transformations.	1
E.2.	Use inductive and deductive reasoning to explore and determine the properties of and relationships among geometric figures.	3
E.3.	Apply trigonometry to problem situations involving triangles and periodic phenomena.	3
F.	MEASUREMENT: Students will understand and demonstrate measurement skills.	2
F.1.	Use measurement tools and units appropriately and recognize limitations in the precision of the measurement tools.	1
F.2.	Derive and use formulas for area, surface area, and volume of many types of figures.	2
G.	PATTERNS, RELATIONS, FUNCTIONS: Students will understand that mathematics is the science of patterns, relationships, and functions.	3
G.1.	Create a graph to represent a real-life situation and draw inferences from it.	3
G.2.	Translate and solve a real-life problem using symbolic language.	3

Table 13
Group Consensus
Maine Mathematics, Mathematics, High School

G.3.	Model phenomena using a variety of functions (linear, quadratic, exponential, trigonometric, etc.).	2
G.4.	Identify a variety of situations explained by the same type of function.	1
H.	ALGEBRA CONCEPTS: Students will understand and apply algebraic concepts.	2
H.1.	Use tables, graphs, and spreadsheets to interpret expressions, equations, and inequalities.	2
H.2.	Investigate concepts of variation by using equations, graphs, and data collection.	3
H.3.	Formulate and solve equations and inequalities.	2
H.4.	Analyze and explain situations using symbolic representations.	2
I.	DISCRETE MATHEMATICS: Students will understand and apply concepts in discrete mathematics.	2
I.1.	Use linear programming to find optimal solutions to a system.	2
I.2.	Use networks to find solutions to problems.	2
I.3.	Apply strategies from game theory to problem-solving situations.	2
I.4.	Use matrices as tools to interpret and solve problems.	2
J.	MATHEMATICAL REASONING: Students will understand and apply concepts of mathematical reasoning.	3
J.1.	Analyze situations where more than one logical conclusion can be drawn from data presented.	3
K.	MATHEMATICAL COMMUNICATION: Students will reflect upon and clarify their understanding of mathematical ideas and relationships.	2
K.1.	Restate, create, and use definitions in mathematics to express understanding, classify figures, and determine the truth of a proposition or argument.	2
K.2.	Read mathematical presentations of topics within the Learning Results with understanding.	1

Appendix B

Data Analysis Tables

Mathematics

SAT Analysis for Maine

Brief Explanation of Data in the Alignment Tables by Column

Table 1

Goals #	Number of standards plus one for a generic standard for each goal.
Standards #	Average number of standards for reviewers. If the number is greater than the actual number in the goal, then at least one reviewer coded an item for the goal/standard but did not find any standard in the goal that corresponded to the item.
Level	The Depth-of-Knowledge level coded by the reviewers for the standards for each goal.
# of standards by Level	The number of standards coded at each level
% w/in std by Level	The percent of standards coded at each level
Hits Mean & SD	Mean and standard deviation number of items reviewers coded as corresponding to goal. The total is the total number of coded hits.
Cat. Conc. Accept.	"Yes" indicates that the goal met the acceptable level for criterion. "Yes" if mean is six or more. "Weak" if mean is five to six. "No" if mean is less than five.

Table 2

	First five columns repeat columns from Table 1.
Level of Item w.r.t. Stand	Mean percent and standard deviation of items coded as "under" the Depth-of-Knowledge level of the corresponding standard, as "at" (the same) the Depth-of-Knowledge level of the corresponding standard, and as "above" the Depth-of-Knowledge level of the corresponding standard.
Depth-of-Know. Consistency Accept.	"Yes" indicates that 50% or more of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding standards. "Weak" indicates that 40% to 50% of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding standards. "No" indicates that less than 40% items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding standards.

Table 3

First five columns repeat columns from Table 1 and 2.

Range of Standards

Standards Hit — Average number and standard deviation of the standards hit coded by reviewers.

% of Total — Average percent and standard deviation of the total standards that had at least one item coded.

Range of Know. Accept. — "Yes" indicates that 50% or more of the standards had at least one coded standard.
"Weak" indicates that 40% to 50% of the standards had at least one coded standard.
"No" indicates that 40% or less of the standards had at least one coded standard.

Balance Index

% Hits in Std/Ttl Hits — Average and standard deviation of the percent of the items hit for a goal of total number of hits (see total under the Hits column).

Index — Average and standard deviation of the Balance Index.

Note: BALANCE INDEX $1 - \left(\sum_{k=1} \left| 1/(O) - I_{(k)}/(H) \right|\right)/2$

Where O = Total number of standards hit for the goal
$I_{(k)}$ = Number of items hit corresponding to standard (k)
H = Total number of items hit for the goal

Bal. of Rep Accept. — "Yes" indicates that the Balance Index was .7 or above (items evenly distributed among standards).
"Weak" indicates that the Balance Index was .6 to .7 (a high percentage of items coded as corresponding to two or three standards).
"No" indicates that the Balance Index was .6 or less (a high percentage of items coded as corresponding to one standard.)

Table 4

Summary if goal met the acceptable level for the four criteria by each goal.

Table 5
Comments made by reviewers on items identified as having a source of challenge issue by item number.

Table 6
The DOK value for each assessment item given by each reviewer. The intraclass correlation for the group of reviewers is given on the last row.

Table 7
All notes made by reviewers on items by item number.

Table 8
The DOK level and standard code assigned by each reviewer for each item.

Table 9
This list for each item all of the standards coded by the eight reviewers as corresponding to the item. Repeat of a standard indicates the number of reviewers who coded that standard as corresponding to the item.

Table 10
This lists for each standard all of the items coded by the eight reviewers as corresponding to the standard. Repeat of an item indicates the number of reviewers who coded the item as corresponding to the standard.

Table 11
This table summarizes the number of reviewers who coded an item as corresponding to a standard. It contains the same information as in Table 10.

Table 12
This table can be used to compare the DOK level of a standard to the average DOK level of the items reviewers assigned to the standard. This table is helpful to identify items with a lower DOK level that should be replaced by an item with a higher DOK level to improve the Depth-of-Knowledge Consistency.

Table 1
Categorical Concurrence Between Standards and Assessment as Rated by Two Reviewers
Maine High School Mathematics
Number of Assessment Items - 54

Standards			Level by Objective			Hits		Cat. Concurr.
Title	Goals #	Objs #	Level	# of objs by Level	% w/in std by Level	Mean	S.D.	
A. NUMBERS AND SENSE	2	3	2	2	100	5.5	1.5	NO
B. COMPUTATION	2	3	2 3	1 1	50 50	8	2	YES
C. DATA ANALYSIS AND STATISTICS	5	5.5	2 3	2 3	40 60	2	2	NO
D. PROBABILITY	2	3	2 3	1 1	50 50	1	0	NO
E. GEOMETRY	3	3.5	1 3	1 2	33 66	12.5	0.5	YES
F. MEASUREMENT	2	2	1 2	1 1	50 50	1	0	NO
G. PATTERNS, RELATIONS, FUNCTIONS	4	5	1 2 3	1 1 2	25 25 50	7.5	2.5	YES
H. ALGEBRA CONCEPTS	4	4.5	2 3	3 1	75 25	23.5	2.5	YES
I. DISCRETE MATHEMATICS	4	4.5	2	4	100	1.5	0.5	NO
J. MATHEMATICAL REASONING	1	2	3	1	100	2	1	NO
K. MATHEMATICAL COMMUNICATION	2	2	1 2	1 1	50 50	0.5	0.5	NO
Total	31	38	1 2 3	4 16 11	12 51 35	65	3	

Table 2

Depth-of-Knowledge Consistency Between Standards and Assessment as Rated by Two Reviewers

Maine High School Mathematics

Number of Assessment Items - 54

Standards			Hits		Level of Item w.r.t. Standard						DOK Consistency
					% Under		% At		% Above		
Title	Goals #	Objs #	M	S.D.	M	S.D.	M	S.D.	M	S.D.	
A. NUMBERS AND SENSE	2	3	5.5	1.5	30	22	53	5	17	24	YES
B. COMPUTATION	2	3	8	2	100	0	0	0	0	0	NO
C. DATA ANALYSIS AND STATISTICS	5	5.5	2	2	100	0	0	0	0	0	NO
D. PROBABILITY	2	3	1	0	100	0	0	0	0	0	NO
E. GEOMETRY	3	3.5	12.5	0.5	56	46	4	5	40	49	WEAK
F. MEASUREMENT	2	2	1	0	0	0	100	0	0	0	YES
G. PATTERNS, RELATIONS, FUNCTIONS	4	5	7.5	2.5	94	12	6	12	0	0	NO
H. ALGEBRA CONCEPTS	4	4.5	23.5	2.5	30	36	64	41	6	11	YES
I. DISCRETE MATHEMATICS	4	4.5	1.5	0.5	0	0	100	0	0	0	YES
J. MATHEMATICAL REASONING	1	2	2	1	100	0	0	0	0	0	NO
K. MATHEMATICAL COMMUNICATION	2	2	0.5	0.5	0	0	0	0	100	0	YES
Total	31	38	65	3	57	45	32	42	10	28	

Table 3
Range-of-Knowledge Correspondence and Balance of Representation Between Standards and Assessment as Rated by Two Reviewers
Maine High School Mathematics
Number of Assessment Items - 54

Standards			Hits		Range of Objectives				Rng. of Know.	Balance Index				Bal. of Represent.
					# Objs Hit		% of Total			% Hits in Std/Ttl Hits		Index		
Title	Goals #	Objs #	Mean	S.D.	Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.	
A. NUMBERS AND SENSE	2	3	5.5	1.5	1.5	0.5	50	17	YES	8	2	0.89	0.11	YES
B. COMPUTATION	2	3	8	2	1.5	0.5	50	17	YES	12	4	0.9	0.1	YES
C. DATA ANALYSIS AND STATISTICS	5	5.5	2	2	1	1	17	17	NO	3	3	0.38	0.38	NO
D. PROBABILITY	2	3	1	0	1	0	33	0	NO	2	0	1	0	YES
E. GEOMETRY	3	3.5	12.5	0.5	2.5	0.5	71	4	YES	19	2	0.58	0.08	NO
F. MEASUREMENT	2	2	1	0	1	0	50	0	YES	2	0	1	0	YES
G. PATTERNS, RELATIONS, FUNCTIONS	4	5	7.5	2.5	3	0	60	0	YES	11	3	0.75	0.02	YES
H. ALGEBRA CONCEPTS	4	4.5	23.5	2.5	4	1	88	12	YES	36	2	0.71	0.04	YES
I. DISCRETE MATHEMATICS	4	4.5	1.5	0.5	1.5	0.5	32	8	NO	2	1	1	0	YES
J. MATHEMATICAL REASONING	1	2	2	1	1.5	0.5	75	25	YES	3	1	0.92	0.08	YES
K. MATHEMATICAL COMMUNICATION	2	2	0.5	0.5	0.5	0.5	25	25	NO	1	1	0.5	0.5	NO
Total	31	38	65	3	1.73	1.04	50	21		9	11	0.78	0.15	

Table 4
Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Two Reviewers
Maine High School Mathematics
Number of Assessment Items - 54

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
A. NUMBERS AND SENSE	NO	YES	YES	YES
B. COMPUTATION	YES	NO	YES	YES
C. DATA ANALYSIS AND STATISTICS	NO	NO	NO	NO
D. PROBABILITY	NO	NO	NO	YES
E. GEOMETRY	YES	WEAK	YES	NO
F. MEASUREMENT	NO	YES	YES	YES
G. PATTERNS, RELATIONS, FUNCTIONS	YES	NO	YES	YES
H. ALGEBRA CONCEPTS	YES	YES	YES	YES
I. DISCRETE MATHEMATICS	NO	YES	NO	YES
J. MATHEMATICAL REASONING	NO	NO	YES	YES
K. MATHEMATICAL COMMUNICATION	NO	YES	NO	NO

Table 5
Source-of-Challenge Issues by Reviewer
Maine High School Mathematics

Item Number	Comments by Reviewer

Table 6
Depth-of-Knowledge Levels by Item and Reviewers
Intraclass Correlation
Maine High School Mathematics

Item	Rater 1	Rater 2
1	1	2
2	2	2
3	1	2
4	2	2
5	2	2
6	1	2
7	2	2
8	2	2
9	2	2
10	2	2
11	2	2
12	2	2
13	2	2
14	2	2
15	2	2
16	2	2
17	2	2
18	3	2
19	1	1
20	2	2
21	2	2
22	1	1
23	2	2
24	2	2
25	2	2
26	2	2
27	1	2
28	2	2
29	2	2
30	1	1
31	2	2
32	2	2
33	3	3
34	1	2
35	2	2
36	2	2
37	3	2
38	2	2
39	1	2
40	1	1

Table 6
Depth-of-Knowledge Levels by Item and Reviewers
Intraclass Correlation
Maine High School Mathematics

Item	Rater 1	Rater 2
41	2	2
42	1	2
43	1	2
44	2	2
45	2	2
46	2	2
47	2	2
48	2	2
49	1	2
50	2	2
51	2	2
52	2	2
53	3	3
54	2	2

Intraclass Correlation: 0.6919
Pairwise Comparison: 0.7963

Table 7
Notes by Reviewer
Maine High School Mathematics

Item Number	Comments by Reviewer
6	I coded G because it is a function item, not just a graphing item.
7	general H because it doesn't deal with problem situations or different representations
8	again, I coded general standards here because all the algebra and relations objectives deal with real-life contexts
9	I'm coding items to the generic B if they are numerical calculation questions (with no estimation or verification) or if they are story problems with only numbers and no algebra involved.
14	You can have a gridded response item that allows for thousands of correct solutions? Whoa.
18	Only by working through this can it be seen to be a DOK 2. Interestingly--by changing the greater than to a less than and changing the central minus to a plus, it becomes a level 3.
21	Item 5.3 requires reasoning, but there is only one logical solution. Therefore, I have used J rather than J.1.
26	Almost an I item, but it doesn't really use counting principles.
29	I said G2 because even if this doesn't require a formula it still requires thinking as though you were using variables.
34	same as my note on #6
38	Asks students to derive a formula. Also could be coded to H.4
38	G2 and H4 are pretty similar.
39	Item 1 Part 1 requires more than one step, but at the high school level this is not very complex for the students.
41	Item 8.3 requires interpreting a table and concept of mean. G.1 is partly a match. C was selected because there is no objective related to average or means.
44	Pythagorean formula
45	Maine standards have at least four objectives that relate to graphs. Items 8.7 asks students to identify a graph that represents a situation. G.1 seems to be the best fit.
46	general H because this requires intuition about manipulating algebraic expressions, but with no contexts associated
48	Items 8.10 is a combination problem. Not included in the Maine standards.
48	General I, because it's a counting problem.
49	8.11 asks students to find a probability, but not of a compound event. Therefore coded item to the generic objective D.
49	Not D1, because it's not a compound event.
50	8.12 requires knowledge of exponents. Exponents are not noted in the Maine standards. Therefore, I put this item under the generic computation standard B.
51	8.13 requires use of proportions. Proportions are not given in the Maine standards.

Table 7
Notes by Reviewer
Maine High School Mathematics

51	tough to code
52	8.14 requires computation of a perimeter. E2 is the best fit.

Table 8

DOK Levels and Objectives Coded by Each Reviewer

Maine High School Mathematics

Item	DOK 0	PObj 0	S1Obj0	S2Obj0	DOK 1	PObj 1	S1Obj1	S2Obj1
1	1	H.3.			2	H.3.		
2	2	E.2.	B.1.		2	E.2.		
3	1	H.4.			2	H.2.	H.4.	
4	2	A.			2	A.		
5	2	E.1.			2	E.1.		
6	1	H.1.			2	H.1.	G.	
7	2	C.	H.3.		2	H.		
8	2	H.3.			2	H.	G.	
9	2	B.1.			2	B.		
10	2	B.1.			2	B.		
11	2	E.2.			2	E.2.		
12	2	B.			2	B.		
13	2	H.3.			2	H.3.		
14	2	H.3.			2	H.1.	H.3.	
15	2	B.			2	H.3.	G.2.	
16	2	E.2.			2	E.2.		
17	2	H.1.	B.		2	H.2.		
18	3	H.3.			2	H.1.		
19	1	A.			1	A.		
20	2	H.3.			2	H.3.		
21	2	J.	C.2.		2	I.3.	J.	
22	1	A.			1	A.		
23	2	I.4.	B.		2	H.1.	B.	
24	2	E.2.			2	E.2.		
25	2	H.1.			2	H.1.		
26	2	E.2.			2	E.		
27	1	H.3.			2	H.3.		
28	2	H.3.			2	H.1.		
29	2	C.			2	G.2.	B.	
30	1	H.4.			1	H.2.		
31	2	E.2.			2	E.2.		
32	2	E.1.			2	A.1.		
33	3	E.2.			3	E.2.		
34	1	H.1.			2	G.	H.1.	
35	2	A.			2	A.		
36	2	H.3.			2	H.3.		
37	3	H.4.	G.1.		2	J.		

Table 8
DOK Levels and Objectives Coded by Each Reviewer
Maine High School Mathematics

Item	DOK0	PObj0	S1Obj0	S2Obj0	DOK1	PObj1	S1Obj1	S2Obj1
38	2	H.3.			2	G.2.	H.4.	
39	1	B.			2	B.		
40	1	E.2.			1	E.2.		
41	2	G.1.	C.		2	H.1.		
42	1	H.3.			2	H.3.	G.2.	
43	1	G.3.			2	H.2.		
44	2	E.2.			2	E.2.		
45	2	G.1.			2	G.1.		
46	2	H.3.			2	H.		
47	2	F.2.			2	G.2.	F.2.	
48	2	B.			2	I.	A.	
49	1	D.			2	D.		
50	2	B.			2	H.3.		
51	2	G.			2	G.		
52	2	E.2.			2	E.2.		
53	3	K.1.	H.3.		3	A.1.		
54	2	E.2.			2	E.2.	J.1.	

Objective Pairwise Comparison: 0.4595
Standard Pairwise Comparison: 0.625

Table 9
Objectives Coded to Each Item by Reviewers
Maine High School Mathematics

Low		Medium		High
2		2.407408		4

1	H.3.	H.3.		
2	B.1.	E.2.	E.2.	
3	H.2.	H.4.	H.4.	
4	A.	A.		
5	E.1.	E.1.		
6	G.	H.1.	H.1.	
7	C.	H.	H.3.	
8	G.	H.	H.3.	
9	B.	B.1.		
10	B.	B.1.		
11	E.2.	E.2.		
12	B.	B.		
13	H.3.	H.3.		
14	H.1.	H.3.	H.3.	
15	B.	G.2.	H.3.	
16	E.2.	E.2.		
17	B.	H.1.	H.2.	
18	H.1.	H.3.		
19	A.	A.		
20	H.3.	H.3.		
21	C.2.	I.3.	J.	J.
22	A.	A.		
23	B.	B.	H.1.	I.4.
24	E.2.	E.2.		
25	H.1.	H.1.		
26	E.	E.2.		
27	H.3.	H.3.		
28	H.1.	H.3.		
29	B.	C.	G.2.	
30	H.2.	H.4.		
31	E.2.	E.2.		
32	A.1.	E.1.		
33	E.2.	E.2.		
34	G.	H.1.	H.1.	
35	A.	A.		
36	H.3.	H.3.		
37	G.1.	H.4.	J.	
38	G.2.	H.3.	H.4.	

Table 9
Objectives Coded to Each Item by Reviewers
Maine High School Mathematics

39	B.	B.	
40	E.2.	E.2.	
41	C.	G.1.	H.1.
42	G.2.	H.3.	H.3.
43	G.3.	H.2.	
44	E.2.	E.2.	
45	G.1.	G.1.	
46	H.	H.3.	
47	F.2.	F.2.	G.2.
48	A.	B.	I.
49	D.	D.	
50	B.	H.3.	
51	G.	G.	
52	E.2.	E.2.	
53	A.1.	H.3.	K.1.
54	E.2.	E.2.	J.1.

Table 10
Items Coded by Reviewers to Each Objective
Maine High School Mathematics

Low		Medium		High
0		3.023256		23

A.	4	4	19	19	22	22	35	35	48											
A.1.	32	53																		
A.2.																				
B.	9	10	12	12	15	17	23	23	29	39	39	48	50							
B.1.	2	9	10																	
B.2.																				
C.	7	29	41																	
C.1.																				
C.2.	21																			
C.3.																				
C.4.																				
C.5.																				
D.	49	49																		
D.1.																				
D.2.																				
E.	26																			
E.1.	5	5	32																	
E.2.	2	2	11	11	16	16	24	24	26	31	31	33	33	40	40	44	44	52	52	54
	54																			
E.3.																				
F.																				
F.1.																				
F.2.	47	47																		
G.	6	8	34	51	51															
G.1.	37	41	45	45																
G.2.	15	29	38	42	47															
G.3.	43																			
G.4.																				
H.	7	8	46																	
H.1.	6	6	14	17	18	23	25	25	28	34	34	41								
H.2.	3	17	30	43																
H.3.	1	1	7	8	13	13	14	14	15	18	20	20	27	27	28	36	36	38	42	42
	46	50	53																	
H.4.	3	3	30	37	38															
I.	48																			
I.1.																				
I.2.																				
I.3.	21																			

Table 10
Items Coded by Reviewers to Each Objective
Maine High School Mathematics

I.4.	23		
J.	21	21	37
J.1.	54		
K.			
K.1.	53		
K.2.			

Table 11

Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)

Maine High School Mathematics

Low		Medium		High
1		1		2

A.	4:2	19:2	22:2	35:2	48:1								
A.1.	32:1	53:1											
A.2.													
B.	9:1	10:1	12:2	15:1	17:1	23:2	29:1	39:2	48:1	50:1			
B.1.	2:1	9:1	10:1										
B.2.													
C.	7:1	29:1	41:1										
C.1.													
C.2.	21:1												
C.3.													
C.4.													
C.5.													
D.	49:2												
D.1.													
D.2.													
E.	26:1												
E.1.	5:2	32:1											
E.2.	2:2	11:2	16:2	24:2	26:1	31:2	33:2	40:2	44:2	52:2	54:2		
E.3.													
F.													
F.1.													
F.2.	47:2												
G.	6:1	8:1	34:1	51:2									
G.1.	37:1	41:1	45:2										
G.2.	15:1	29:1	38:1	42:1	47:1								
G.3.	43:1												
G.4.													
H.	7:1	8:1	46:1										
H.1.	6:2	14:1	17:1	18:1	23:1	25:2	28:1	34:2	41:1				
H.2.	3:1	17:1	30:1	43:1									
H.3.	1:2	7:1	8:1	13:2	14:2	15:1	18:1	20:2	27:2	28:1	36:2	38:1	42:2
	46:1	50:1	53:1										
H.4.	3:2	30:1	37:1	38:1									
I.	48:1												
I.1.													
I.2.													
I.3.	21:1												
I.4.	23:1												

Table 11
Number of Reviewers Coding an Item by Objective (Item Number: Number of Reviewers)
Maine High School Mathematics

J.	21:2	37:1
J.1.	54:1	
K.		
K.1.	53:1	
K.2.		

Table 12
Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])
Maine High School Mathematics

Low DOK		Matched DOK		High DOK
1		1		2

A. [2]:	4:2[2]	19:2 [1]	22:2 [1]	35:2 [2]	48:1 [2]						
A.1. [2]:	32:1 [2]	53:1 [3]									
A.2. [2]:											
B. [3]:	9:1[2]	10:1 [2]	12:2 [2]	15:1 [2]	17:1 [2]	23:2 [2]	29:1 [2]	39:2 [1.5]	48:1 [2]	50:1 [2]	
B.1. [3]:	2:1[2]	9:1[2]	10:1 [2]								
B.2. [2]:											
C. [3]:	7:1[2]	29:1 [2]	41:1 [2]								
C.1. [3]:											
C.2. [3]:	21:1 [2]										
C.3. [2]:											
C.4. [2]:											
C.5. [3]:											
D. [3]:	49:2 [1.5]										
D.1. [2]:											
D.2. [3]:											
E. [3]:	26:1 [2]										
E.1. [1]:	5:2[2]	32:1 [2]									
E.2. [3]:	2:2[2]	11:2 [2]	16:2 [2]	24:2 [2]	26:1 [2]	31:2 [2]	33:2 [3]	40:2 [1]	44:2 [2]	52:2 [2]	54:2 [2]
E.3.											

Table 12
Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])
Maine High School Mathematics

[3]:													
F. [2]:													
F.1. [1]:													
F.2. [2]:	47:2 [2]												
G. [3]:	6:1[2]	8:1[2]	34:1 [2]	51:2 [2]									
G.1. [3]:	37:1 [3]	41:1 [2]	45:2 [2]										
G.2. [3]:	15:1 [2]	29:1 [2]	38:1 [2]	42:1 [2]	47:1 [2]								
G.3. [2]:	43:1 [1]												
G.4. [1]:													
H. [2]:	7:1[2]	8:1[2]	46:1 [2]										
H.1. [2]:	6:2[1.5]	14:1 [2]	17:1 [2]	18:1 [2]	23:1 [2]	25:2 [2]	28:1 [2]	34:2 [1.5]	41:1 [2]				
H.2. [3]:	3:1[2]	17:1 [2]	30:1 [1]	43:1 [2]									
H.3. [2]:	1:2[1.5]	7:1[2]	8:1[2]	13:2 [2]	14:2 [2]	15:1 [2]	18:1 [3]	20:2 [2]	27:2 [1.5]	28:1 [2]	36:2 [2]	38:1 [2]	42:2 [1.5]
	46:1 [2]	50:1 [2]	53:1 [3]										
H.4. [2]:	3:2[1.5]	30:1 [1]	37:1 [3]	38:1 [2]									
I. [2]:	48:1 [2]												
I.1. [2]:													
I.2. [2]:													
I.3. [2]:	21:1 [2]												
I.4. [2]:	23:1 [2]												
J. [3]:	21:2 [2]	37:1 [2]											
J.1.	54:1												

Table 12
Assessment Item DOK vs Consensus DOK (Item Number: Number of Reviewers [Average DOK])
Maine High School Mathematics

[3]:	[2]
K. [2]:	
K.1. [2]:	53:1 [3]
K.2. [1]:	